Bridging Data Gaps: Harnessing Semantic Associations for Knowledge Discovery in Colonial Heritage

Sarah Binta Alam Shoilee*
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

Victor de Boer
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

Annastiina Ahola
Aalto University
Espoo, Finland

Heikki Rantala
Aalto University
Espoo, Finland

Eero Hyvönen
Aalto University
Espoo, Finland

Jacco van Ossenbruggen
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

Susan Legene
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

ABSTRACT

Cultural heritage data, particularly from colonial contexts, frequently presents an incomplete and biased view, reflecting historical institutional priorities more than contemporary knowledge requirements. Consequently, knowledge graphs derived from these records often contain incomplete, fragmented, and skewed data, including absent attributes or values, missing semantic links, and under-represented perspectives. This work addresses the challenge of knowledge discovery under such limitations, presenting a real-world case study on the provenance research of colonial cultural heritage. We present a task-aware design method for building a tool to facilitate this process. The design approach of this application is rooted in a Knowledge Discovery in Database (KDD) framework and is particularly novel due to its formalisation and operationalisation of three distinct types of semantic association: explicit, abstract, and implicit. These semantic associations, grounded in domain interpretation, are crucial for bridging data gaps where user information needs cannot be directly met by existing data. We further demonstrate how these associations can be effectively communicated through user interface components, enabling users to infer new knowledge. We evaluated the resultant application through a user study among domain experts to assess its efficacy. The evaluation confirms the effectiveness of the tool in enabling new knowledge discovery and reveals opportunities to improve the representation of the underlying data, as users could successfully infer insights even when information was missing or poorly captured in the original data sets.

CCS CONCEPTS

• **Information systems → Expert search**; • **Human-centered computing → Information visualization**.

*Corresponding author.

K-CAP 2025, December 10–12, 2025, Dayton, Ohio, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Exploratory Knowledge Discovery, Linked Data, Cultural Heritage, Provenance Research

ACM Reference Format:
Sarah Binta Alam Shoilee, Victor de Boer, Annastiina Ahola, Heikki Rantala, Eero Hyvönen, Jacco van Ossenbruggen, and Susan Legene. 2025. Bridging Data Gaps: Harnessing Semantic Associations for Knowledge Discovery in Colonial Heritage. In *Proceedings of The Thirteenth International Conference on Knowledge Capture (K-CAP 2025).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Provenance research is a crucial field in cultural heritage studies, focusing on tracing the origins, ownership history, and movement of objects across time and space [25]. It provides essential insights into the ethical, legal, and historical contexts of heritage objects' collections, particularly in cases of contested ownership, colonial acquisitions, and restitution claims [20]. By reconstructing an object's past, provenance research helps museums, scholars, and policymakers make informed decisions about collections, ensuring transparency, accountability, and historical justice. The very nature of this pursuit, however, presents a compelling challenge for the field of knowledge discovery (KD) due to its pressing need for large-scale, structured analysis across historically constructed datasets. These datasets were often originally created for administrative or archival purposes, not for investigative exploration of objects' provenance, and thus contain missing attributes, incomplete values, and fragmented or implicit links between entities. Moreover, the data reflect the priorities, perspectives, and limitations of those who recorded it over the centuries, often resulting in under-representation of certain viewpoints and a partial or biased account of historical events [28]. This makes "new" knowledge discovery from the existing database for Provenance Research a non-trivial task, as it not only requires query answering or pattern recognition, but also requires exploring data quality, implicit semantic associations within the data, and user or domain interpretation for knowledge inference, so that the KD system can adapt to gaps and uncertainties of the data.

Knowledge Discovery is a key task advanced by Linked Open Data (LOD) [13], especially as its adaptation grows in critical domains, such as cultural heritage [29]. Beyond information retrieval,

these studies demonstrate increasing demands for tools that support open-ended exploration and insight generation. The inherently connected structure of Linked Data encourages the "Follow-Your-Nose" navigation style, supporting discovery through traversal [16]. Previous studies [5, 10, 30] have shown that LOD promotes serendipitous outcomes in exploratory search, revealing unexpected but meaningful connections. Exploratory search or Exploratory Data Analysis (EDA) [26] is particularly valuable for users with evolving information needs and unfamiliar with underlying data structure, helping them discover patterns through iterative visualisation and summarisation. Building on the broader framework of the Knowledge Discovery in Databases (KDD) process [7], which includes data preparation, pattern search, knowledge evaluation, and refinement, this study focusses on KD through the lens of EDA, referring to it as Exploratory Knowledge Discovery (EKD) [3].

In this paper, we tackle the challenge of enabling knowledge discovery from *incomplete*, *semantically fragmented* and *biased* knowledge graphs for colonial heritage object provenance research. Through our case-study, we observe that the available knowledge graph (KG) lacks key attributes, contains missing or misplaced values, and fragmented links that hinder direct querying and automated reasoning for the domain's information need. To address this, we investigate: *How can interesting knowledge discovery be supported in Linked Data environments when data gaps and semantic fragmentation prevent direct answers to provenance-related questions?* To answer this question, we contribute a three-part methodologically grounded solution: (1) we design and implement a web application based on the KDD framework that supports EKD for provenance research, accommodating incomplete and fragmented historical data from our case study. (2) Central to this application design is our novel contribution: the formal definition and operationalisation of three types of semantic associations — explicit, abstract, and implicit — which are grounded in domain interpretation and designed to bridge data gaps. These associations systematically model how human experts make sense of incomplete data. (3) We further evaluate the resultant system through user study to assess its effectiveness in facilitating knowledge discovery and uncovering hidden knowledge within the given context.

This paper builds on two earlier short papers [21, 22] in which we introduced the application and demonstrated its functionality. In the current paper, we extend those works by formally presenting the methodological foregrounding and design choices that guided the development of the application, as well as the design and execution of a user study to evaluate its effectiveness.

2 BACKGROUND AND RELATED WORK

Knowledge Discovery in Databases (KDD) [9] is defined as a multi-step process for identifying valid, novel, potentially useful, and ultimately understandable patterns in data, as outlined by Fayyad et al. [7]. A key distinction between knowledge discovery (KD) and data analysis (one step in KDD) lies in the notion of interestingness, a measure that reflects the value of discovered patterns through an analytical algorithm. EKD [3] extends the KDD process by shifting the focus from algorithmic pattern extraction to open-ended user-driven exploration, aligning with the tradition of 'exploratory data analysis' [26]. While KDD focusses on mining algorithmic patterns, EKD foregrounds human-led discovery via visual interfaces. Consequently, we use the term 'findings' rather than 'patterns' to describe the outcomes of the EKD – observations or associations formed through exploration that contribute to contextual understanding [19]. Unlike patterns, findings do not need to be generalised or formalised rules; instead, they capture the nuanced, contextual, and often serendipitous nature of knowledge discovery that EKD systems are designed to support.

We hypothesise that EKD is an appropriate methodology for addressing the multifaceted challenges inherent in the incomplete, biased, and fragmented nature of the data. The core strength of EKD lies in its ability to facilitate robust and trustworthy KD by actively involving the human user in an iterative and interpretive process, which is crucial when users have limited understanding of the data in advance [26]. This approach enables users to navigate inherent data uncertainties, iteratively form hypotheses, and leverage their domain expertise to bridge gaps and validate insights, thus leading to more reliable conclusions. Furthermore, EKD directly supports the human interpretive processes essential for deciphering the implicit meanings and relationships within the fragmented dataset, offering tools and interfaces that scaffold the conversion of tacit human understanding into new knowledge [3], making reliable and comprehensive discoveries possible even from demonstrably incomplete knowledge graphs. Therefore, the justification for adapting EKD as a knowledge discovery process in the given context.

Given the broad body of work on Exploratory Knowledge Discovery (EKD), several previous studies have influenced our research. The open-world and interconnected nature of Linked Data makes it a natural fit for EKD tasks. However, SPARQL query alone may not be well suited for inexperienced user in exploratory tasks, various approaches have been proposed to augment semantic portals, most commonly through faceted search interfaces [11, 27], or visual exploration techniques [4]. Multiple studies have also explored the potential of Linked Data to support serendipitous knowledge discovery [5, 10, 16, 30]. Despite these advancements, there is a lack of systematic methodologies and tools for knowledge discovery from knowledge graphs with incomplete, fragmented, and skewed data. More specifically, existing work does not address how domain-specific requirements can be structurally incorporated even when user information needs cannot be directly queried given the scope of the dataset. In this paper, we present a tool design approach to address this gap. Using colonial object provenance research from a real world case study, we demonstrate how our designed system supports EKD and provides concrete examples to facilitate insight generation, uncover hidden relationships, and support complex research tasks in data-fragmented domains.

This work uses Sampo-UI [15, 18], a well-established and reusable framework and part of the Sampo Model [12] to build semantic portals for cultural heritage, as shown by more than 20 in-use applications[1], such as BookSampo [2] and ParliamentSampo [14]. These systems underscore the benefits of semantic data reuse, shared ontologies, and faceted exploration integrated seamlessly with data analyses. Our work expands Sampo-UI's potential by addressing the challenge of bridging data gaps in KG by capturing and operationalising implicitly understood meanings and relationships by

[1] See https://seco.cs.aalto.fi/applications/sampo/ for further information and links.

domain experts in historical data. This work formally defines and implements various types of semantic association (explicit, abstract, and implicit). These associations cannot always be directly identified from the data, but instead necessitate domain interpretation for their identification and utilisation.

3 METHODOLOGY

In this section, we present a design approach for developing an EKD application in Linked Data environments, particularly suited to datasets characterised by missing attributes or values, fragmented semantic links, and skewed information. Assuming that the Linked (Open) Data is already available, we focus on the downstream discovery pipeline, which we structure into five stages that are subsequently compared to the KDD steps of Fayyad et al. [8].

Stage-1: User Requirement Analysis. The first steps in the KDD process involve understanding the application domain, prior knowledge, and user goals. To integrate domain-specific knowledge and user goals into the KD process, we propose repurposing competency questions (CQs) [31], which are traditionally used to define ontology requirements and guide system design. This reuse is grounded in the value of CQs for articulating core domain concerns and identifying key entities and relationships in KG. In cases where information needs cannot be met directly by answering CQs due to missing attributes, information gaps, or fragmented links, we argue that CQs can yet serve as a navigational guideline within the exploratory process. They provide a foundation for exploration that can be further supported by interactive user interfaces that enable iterative refinement, branching paths, and interpretive reasoning for users. This interplay between structured inquiry and flexible navigation not only facilitates targeted exploration but also allows for the emergence of serendipitous and previously unforeseen insights. Even in EKD settings, grounding discovery in the domain context is essential [3], as user goals and knowledge shape meaningful exploration. Domain-informed design choices, such as specific semantic filters and entity associations, help ensure that the system supports both targeted inquiry and serendipitous insight.

Stage-2: Identification of Entity Types and Semantic Associations of Interest. At this stage, we consider identifying key entity types that reflect domain-specific requirements derived from CQs. These include both the user's inquiry focus – referred to as *target entities* – and the *supporting entities* that provide the necessary context or information to address the CQ. For example, in the question "What is the production date of Object X?", the production date represents the target entity, while the object (Object X) serves as a supporting entity. Together, these define the relevant *entity types of interest* for exploration. To assess whether the KG can support such queries, we recommend inspecting the available Linked Data using resources such as RDF schemas, SHACL shapes, or graph statistics – whichever are accessible – to gain a structural understanding of the data and discard entity types of interest if the Knowlege Graph (KG) do not have such entities. These operations align with the steps of *data selection* and *pre-processing and cleaning* in KDD.

Building on the identification of relevant entity types of interest, our methodology systematically uncovers interesting semantic associations and their corresponding domain questions. As detailed in Process 1, this process takes the previously identified *entity types of interest*, E, as input. Then it generates all unique pairwise combinations of these entity types. For each pair, a natural language question is manually constructed based on semantic coherence given the data. Crucially, these questions are then presented to a domain expert, who assesses their relevance within the specific domain context. If the expert confirms a question's relevance, the corresponding entity pair is added to a set of *interesting semantic associations*, and the question itself is added to a set of interesting *domain questions*. This expert-validated process ensures that the identified semantic associations and questions are not merely queriable, but useful for guiding deeper EKD within the domain.

Process 1 Identifying Interesting Semantic Associations and Domain Questions

Input: Entity types of interest E
Output: Set of Interesting Semantic Associations A, Set of Interesting Domain Questions A

$P \leftarrow \{(e_1, e_2) \mid e_1, e_2 \in E\}$ {All 2-element combinations}
$A \leftarrow \emptyset$
$D \leftarrow \emptyset$
for all $(e_1, e_2) \in P$ **do**
  Generate natural language question Q for (e_1, e_2)
  if Q is structurally and semantically valid **then**
    Ask domain expert if Q is relavent
    if Expert confirms **then**
      $A \leftarrow A \cup \{(e_1, e_2)\}$
      $D \leftarrow D \cup Q$
    end if
  end if
end for
return A and D

Stage-3: Mapping Semantic Association to UI Components. Our approach categorises semantic associations into three distinct types, based on the interpretation required to answer their corresponding domain questions. (1) *Explicit associations* are those semantically defined by the schema of the knowledge graph, which allows them to be answered by SPARQL queries without further domain interpretation. E.g., <objectX, producedBy, personY>. Our pipeline mainly uses these associations for faceted search and filtering mechanisms [27], empowering users to refine results based on explicitly modelled relationships. (2) *Abstract associations* represent high-level relationships that are not semantically represented in the data schema but can be structurally derived by domain interpretation. For example, given <personX, participatedIn, EventA> and <personY, participatedIn, EventA>, based on domain interpretation, the semantic association can be drawn <personX, relatedTo, personY>. These associations are computed at runtime using SPARQL queries based on domain-informed graph patterns. These associations are useful for intuitive exploration through hyperlinks, fostering a "follow-your-nose" style of exploration [32]. Finally, (3) *Implicit associations* are the most challenging, as they are not structurally

represented in the knowledge graph and therefore cannot be retrieved via SPARQL. Instead, they are inferred by users through analysis of visual patterns and contextual cues. For example, "Are there any patterns in object acquisition from *PlaceX* by *ActorY* around *TimeZ*, suggesting connection to *EventA*?". To make these possible, our approach integrates exploratory analytics tools [26] such as timelines, maps, and network diagrams, which enable users to perceive associations based on visual proximity or co-occurrence. This conceptual mapping largely guides our interface design and operationalise these different types of semantic association.

This identification and communication of semantic associations are analogous to the data mining and pattern interpretation stages of the traditional KDD process, which are modified to support EKD and meet data requirements. User interaction with all these different types of semantic associations is vital for KD in incomplete, biased, and fragmented knowledge graphs, when CQs cannot be answered directly. The exhaustive pairwise construction of semantic association and consequently domain questions from the previous stage enables a more holistic opportunity for KD, surfacing explorative and interpretive insights beyond initial CQs. By formalising these diverse association types, we provide varied interaction strategies that allow users to leverage existing data more comprehensively, move from direct retrieval to expert interpretation, and thereby reveal hidden potential in incomplete datasets.

Stage-4: Evaluation. To evaluate effectiveness of the designed EKD system, our pipeline incorporates a user study to determine if the findings discovered through the designed system are "interesting" for the intended end user, ultimately resulting in KD. This evaluation is grounded in understanding that *interestingness* is considered a core measure of KD [24]. Since interestingness is inherently a subjective measure, user feedback becomes essential for assessing whether the system achieves its goal of enabling KD. During the evaluation phase, the approach addresses the following questions: (1) Do users discover findings that they consider interesting through the portal? (2) If so, what particular features facilitate such discoveries? (3) Do users' perceptions of "interestingness" align with our operational definition? Finally, (4) How can the system be further refined to enhance its effectiveness?

Our approach to assess user-perceived interestingness during exploration uses the free recall method [17] along with the think-aloud protocol [1]. Participants are encouraged to navigate the designed system autonomously, driven by their own curiosity, while continuously verbalising their thoughts and noting any findings that they deem interesting. To operationalise "interestingness", we draw upon the definition by Silberschatz et al. [24], focussing on two key dimensions: (1) usefulness, where a finding enables meaningful action (actionability), and (2) unexpectedness, when a finding is surprising. To align user perceptions with this definition, participants were asked to evaluate each listed findings by rating their agreement with given statements on usefulness and unexpectedness using a 5-point Likert scale (ranging from -2 for strongly disagree to +2 for strongly agree). A positive score in either statement signifies a successful knowledge discovery. Beyond these structured evaluations, we also gather feedback through qualitative observation of user interactions, allowing us to pinpoint usability issues, areas of confusion, or misinterpretations. Following the exploration session, participants provide additional structured and open-ended feedback, offering suggestions and recommendations. This comprehensive input is invaluable for guiding iterative improvements in the system's usability, reliability, and its overall support for EKD.

4 USE-CASE IMPLEMENTATION

Provenance research is a multidisciplinary effort focused on tracing the ownership history and historical context of cultural objects [25]. This task becomes particularly challenging in the context of collections from the colonial era, where the data available in the museum database are often fragmented, incomplete, and biased [28]. In such contexts, the overarching task, which is reconstructing objects' biography, becomes exploratory. Users, such as curators, historians, and provenance researchers often find that the questions they want to ask to the databases system cannot be answered through straightforward queries given the shortcoming of the data. Even when relevant information exists, it is often hidden within implicit relationships or scattered across the database. As a result, users often find themselves engaged in a process of sense making: mining, connecting, and interpreting data to uncover hidden associations.

Table 1: Key entities, representative classes and instance counts in the dataset

Entity Types	CIDOC-CRM Classes	Number of Instances
Object	`E22_Human-Made_Object`	1,039,164
Production	`E12_Production`	658,715
Production Actor	`E21_Person, E74_Group, E39_Actor`	4,866
Production Place	-	5,730
Production Time-period	`E52_Time-Span`	997,786
Historical Event	`E5_Event`	328
Acquisition Event	`E8_Acquisition`	1,161,521
Acquisition Time-period	`E52_Time-Span`	662,273
Transfer of Custody Event	`E10_Transfer_of_Custody`	174,967
Provenance Actor	`E21_Person, E74_Group, E39_Actor`	17,099

In this section, we implement the EKD application following Section 3 for provenance research case study and using existing datasets. We use data from the Wereldmuseum[2], initially recorded in The Museum System (TMS)[3], and later Linked Data published through the Colonial Collections Data Hub[4]. TMS data, often converted from physical index cards, are curated by museum experts. So, data may refer date back centuries or be updated with new research insights, reflecting personal or institutional viewpoints. TMS, supported by a relational database, is limited in comprehensively documenting provenance events, such as transfer of custody or acquisition details, which is critical for provenance research. The Colonial Collections Data Hub has published Wereldmuseum's TMS data as Linked Data with a public SPARQL endpoint[5], enhancing access and structure of object metadata through CIDOC-CRM [6] alignment. The Linked Data supports structured provenance metadata and includes events like `E8_Acquisition` and `E10_Transfer_of_Custody`. However, the underlying TMS database lacks detailed event-centric metadata documentation, limiting representation of these provenance events and resulting in sparse provenance information (see Table 1).

[2]Wereldmuseum webpages: https://\{amsterdam/leiden.rotterdam\}.wereldmuseum.nl
[3]TMS: https://www.gallerysystems.com/solutions/collections-management/
[4]Colonial Collection Hub data portal: https://data.colonialcollections.nl
[5]Wereldmuseum Linked Data endpoint: https://api.colonialcollections.nl/datasets/nmvw/collection-archives/sparql

4.1 User Requirement Analysis

We gathered user requirements to understand the information needs of provenance research on heritage objects as a form of competency questions (CQs). These CQs were derived from a previous study [23], which was developed through interviews with five museum professionals. Among them were three provenance researchers and two postdoctoral researchers with backgrounds in ethnographic collections. All interviewees had deep domain expertise in different geographic or historical contexts, ranging from East Africa, Central and Southern Africa, and Asia to missionary collections and human remains. These competency questions serve as the basis for further exploratory analysis. The CQs are presented in Table 2, identifying the *entity types of interest*.

Table 2: Competency Questions from previous paper with *entity type of interest* in bold.

CQ-1	Which **persons** were involved in the provenance of this **object**?
CQ-2	Which **objects** are collected by **person A**?
CQ-3	Is there a relationship between **person A** and **person B**?
CQ-4	Which **objects** were collected in this **geographical location**?
CQ-5	Which **objects** were collected during this **(historical) event**?
CQ-6	Which **objects** were collected in this **geographical location** during this **time period**?
CQ-7	Which source states this statement?
CQ-8	Who or which institution conducted this research?
CQ-9	Which is the latest version of the provenance research?

4.2 Identification of Entity Types and Semantic Associations of Interest

By examining the graph statistics of the dataset (Table 1) and the repetition of entity types in the competency questions, we identified five primary entity types of interest. These include: (1) **Object**, instances of E22_Human-Made_Object, which denotes heritage objects in the museum and serves as the central node in the data; (2) **Actor**, described using E39_Actor, E21_Person, and E74_Group, encompassing individuals, institutions, or groups involved in the creation, acquisition, or transfer of custody of objects; (3) **Event**, referring to historical events influencing the object's provenance expressed as E5_Event; (4) **Time**, which contextualises object creation and acquisition event, expressed as E52_time-span; and (5) **Place**, representing geographical locations linked to the object's biography. Entity types, i.e., `E8_Acquisition` and `E10_Transfer_of_Custody`, though frequent in the data and vital for modelling object provenance, were not directly mentioned in user-defined competency questions and lacked relevant attribute values. Therefore, we do not classify them as entity types of interest from a user-driven exploration perspective. Note that, we do not consider CQ7-9, as the current dataset does not support that.

After identifying five core entity types of interest, $E = \{Object, Actor, Historical\ Event, Time, Place\}$, we explored all subsets of two entities to establish binary associations as described Section 3 stage 2. This resulted in ten unique subsets and five additional same-type associations (e.g., Object-Object), totaling 15 combinations for semantic associations. For each pair, we developed domain-relevant questions based on potential associations and validated by a domain expert. We disregarded few combinations ({place, place}, {time, time}) due to their inability to meaningfully create any domain-specific question. Given the data scope in Table 1, not all CQs can be directly addressed. For instance, events like `E10_Transfer_of_Custody` and `E8_Acquisition` lack a place connection, preventing direct answers to CQ-4 or CQ-6. However, examining the {place, object} pair can offer indirect insights. Here, 'place' refers to the production location or *Place of Origin* of the object, not its collection spot. Identifying such implicit associations is important. Some domain-questions directly answers to CQs and some help address them. This mapping of domain-questions to CQs is shown in Table 3. We determine which type of semantic association, explicit, abstract, or implicit, answers the questions based on the data structure.

4.3 Mapping Semantic Association to UI Components

The data, though modelled using CIDOC-CRM for event-centric representation, come from a museum collection system focused on object cataloguing. Consequently, the resultant knowledge graph is predominantly object-centred. Provenance events such as production, acquisition, and transfer of custody are linked to objects through relevant CIDOC-CRM properties, and these events in turn are associated with various attributes such as places, times, and involved actors. Historical events are expressed with a property chain `crm:P141i_was_assigned_by / crm:P141_assigned` connecting to objects. Since most entities are directly connected to objects or linked via production and provenance events, many of the associations involving objects has *explicit association*. Examples of these associations include: {actor, object}, {event, object}, {time, object}, and {place, object}. To support this, we simplify certain property paths by introducing direct links. For example, the path `crm:P108i_was_produced_by / crm:P14_carried_out_by`, which indicates the maker of an object, was replaced with direct property `ex:maker` which results which ultimately alter the data model and made accessible via a dedicated SPARQL endpoint[6][7].

For certain entity combinations, i.e., {actor, event}, {actor, time}, {actor, place}, {event, time}, and {event, place}, {actor, actor}, there is no semantic path between these entities based on existing ontology. However, domain-interpreted connections can be inferred through intermediate entities, typically via the object. For example, if a historical event attributed to object collection and an actor is linked to that object (e.g., as collector), a <relatedTo> relationship or *abstract association* between the actor and the historical event can be drawn based on domain reasoning.

For entity combinations where no direct or logically inferred path can be established through SPARQL queries, but semantically interesting associations may emerge through underlying data patterns, are considered for *implicit association*. For example, {time, place} combination, there might be a trend in the place of origin and the object acquisition date that we cannot establish through graph patterns, but using visual analytics. Same goes for {place, event} association from the data pattern relation can be drawn.

Implementation of the semantic portal Following these decision, we implemented PM-Sampo, a web application built on

[6]Data endpoint: http://ldf.fi/pm-sampo/sparql
[7]Data and its documentation: https://github.com/Shoilee/PM-SampoDataManager

Table 3: Overview of semantic associations, corresponding domain questions (with referring CQ), and PM-Sampo UI components supporting those. The related UI components are implemented under the entity perspective that is mentioned in bold.

Entity Pair	Type of Semantic Association	Domain-Questions	User Interface (UI) Components
{actor, **object**}	Explicit	Which objects were acquired from actor A? **(CQ-2)**	Faceted search in object perspective.
{actor, **object**}	Abstract	Which actors were involved in the provenance of this object? **(CQ-1)**	Object instance page, provenance tab.
{**actor**, object}	Explicit	Which objects are connected with the same actor? **(CQ-2)**	Actor instance page, object tab OR Filter by actor in objects perspective.
{event, **object**}	Explicit	Which objects were collected during a given historical event? **(CQ-5)**	Faceted Search in objects perspective.
{place, **object**}	Explicit	Which objects were collected from a given place of origin? **(CQ-4)**	Map visualisation of place and objects.
{time, **object**}	Explicit	Which objects were acquired during a given year? **(CQ-6)**	Timeline with objects and acquisition dates.
{**actor**, event}	Abstract	Which historical events is this actor attributed to and with which role? **(CQ-5)**	Event tab in actor instance page with roles.
{actor, **event**}	Abstract	Which actor are related to objects attributed to a historical event?	Instance tab of actor in event instance page.
{actor, time}	Abstract & Implicit	What is the common acquisition time for **objects** collected by this actor? **(CQ-2)**	Timeline of **(objects')** acquisition filtered by actor.
{actor, time}	Abstract & Implicit	Which actor is associated with the highest number of acquisitions in years with significant acquisition activity? **(CQ-2)**	Actor vs **(objects')** acquisition timeline.
{actor, place}	Abstract & Implicit	What are the places this actor has collected objects from?	Map of **objects'** collection locations by actor.
{time, event}	Abstract & Implicit	What are the common acquisition times attributed to a historical event? **(CQ-5)**	**Objects** acquisition timelines filtered by event.
{place, event}	Implicit	What are the common places of origin for objects attributed to a given historical event? **(CQ-4)**	Heatmap showing **object**-place of origin correlations filtered by event.
{time, place}	Implicit	How did the geographical patterns of object collection change over time? **(CQ-6)**	Animated map with **Objects'** place of origin and timeline
{**actor**, actor}	Abstract & Implicit	Is there any relation between actor A and actor B? **(CQ-3)**	Actor-actor network graph.

the Sampo-UI framework [15, 18][8] which offers a strong base for JavaScript-based Linked Data applications with minimal customisation. The framework allows developers to reuse components such as faceted search, data tables, and visualisations through a specification-based configuration. Following the Sampo model [12], our portal uses a SPARQL endpoint for data integration, which supports modular development. In Sampo-UI, each entity type of interest can have *perspective*, which is a faceted semantic search view for filtering and exploring instances of the perspective class(es). Initially, we considered 5 different perspectives: Object, Actor, Historical Event, Time, and Place. However, Time and Place perspectives were excluded because they are mainly attributes of other entities. Each perspective involves two main types of UI components: (1) Faceted search components to filter instances of the perspective class(es) using explicit semantic association. Facets include attributes and other related entities with dynamic hit counts. (2) Visual components for filtered results through tools such as maps, timelines, networks, and tables for semantic exploration.

For the *explicit and abstract association* depending on the selected application perspective, one entity is treated as the target entities, and the others are implemented as faceted filters or supporting entities. For *abstract association*, these connection also visualised on instance pages of individual perspectives (e.g., Actor and Historical Events) through "Related {*Entity*}" tab, improving user access to inferred relationships. For all types of association, but mostly for *implicit association* we try to surface their association through the implementation of visual analytic tools. For instance, the {time, place} association is communicated through animations to highlight spatio-temporal patterns. Detailed mappings examples are provided in Table 3. Furthermore, data enrichment processes were applied, such as, GeoNames data extraction through federated query, to allow geospatial mapping by associating latitude and longitude coordinates with relevant production places. An online demonstrator of the PM-Sampo is available on-line[9]; and the source code for the demonstrator has been released on GitHub[10].

[8]Sampo-UI home: https://seco.cs.aalto.fi/tools/sampo-ui/; Github: https://github.com/SemanticComputing/sampo-ui

[9]PM-Sampo online demonstrator: https://pmsampo.demo.seco.cs.aalto.fi/

[10]PM-Sampo source code: https://github.com/Shoilee/PM-Sampo

4.4 Evaluation

User-study Our user study assessed the knowledge discovery potential of PM-Sampo with five cultural heritage professionals: three provenance scholars and two data registrars. These participants had 1-7 years of experience with museum databases and 3-10+ years in provenance research, and were familiar with the Wereldmuseum collection and TMS data challenges. We analyse observational data and survey responses to address four key evaluation questions (Section 3, stage-4). The study used a subset of 26,180 Wereldmuseum objects' metadata, choosing objects those documented as related to historical events for their higher data quality. A pilot session refined the user study protocol, with all materials publicly available[11].

The study began with an overview, informed consent, and a brief demographic questionnaire. User interaction with PM-Sampo was evaluated through two perspectives: "objects" and "actors". Each perspective involved a video demonstration (15 minutes for objects, 7 minutes for actors) followed by 10 minutes hands-on exploration. Participants were asked to identify and list five "interesting findings" — new facts, patterns, connections, or insights — following their own research interests or curiosity, starting with specific filter facets. Qualitative observations during the session captured navigation, moments of curiosity, and noted any inconsistencies, bugs, or features used. This process were repeated for two perspectives.

The study concluded with a structured survey on: (1) Interestingness Assessment: Participants rated the usefulness and unexpectedness of their findings (1-10) on a Likert scale (-2 to +2). (2) Tool Evaluation: Participants rated their agreement with four statements on PM-Sampo's overall effectiveness in supporting exploratory knowledge discovery (EKD), using the same Likert scale: (Q1) *I would consider using PM-Sampo in my own research for exploring heritage object provenance.* (Q2) *I think PM-Sampo has potential to uncover relationships or insights I might not have discovered otherwise.* (Q3) *I plan to use PM-Sampo in my future research to support knowledge discovery.* (Q4) *PM-Sampo enables exploratory analysis that aligns with my research interests.* Finally, participants provided additional comments, suggestions, and listed the three most useful features.

[11]User-study protocol, materials and results: https://doi.org/10.5281/zenodo.15423716

Results *Do users discover findings that they consider interesting through the portal?* All participants overall documented 36 unique findings, averaging about seven per participant, demonstrating the portal's ability to facilitate interesting discoveries in under 20 minutes. Observations revealed that the portal triggers curiosity, leading to semantic exploration. For example, a participant (P1) noticed a surge of objects from Morocco, as illustrated in the *Place of Origin map*. This initial burst of interest, triggered by an implicit association, prompted the user to use the facet filter for "Place of Origin: Morocco" to find the main contributor: J.E. (Josephine) Powell, thus hinting an {actor, place} connection. This indicates that the portal aids users in transitioning from visual insights to structured queries. More such observations are available in the Zenodo repository.

What particular features facilitate such discoveries? All participants began their exploration with the *Production Places* view of *object* perspective. This was primarily because it allowed them to visually locate geographic areas of interest and to see how many objects the museum holds from that region. Although the place of origin was recorded using regional labels, the map visualisation grouped objects by precise locations and provided count-based indicators. These visual cues often triggered participants' curiosity, prompting deeper investigation. Another widely appreciated feature was the "Connected Historical Event" connection. Participants noted that it helped them uncover high-level or abstract associations between actors and historical events, revealing underlying contextual links that were not immediately obvious. In addition, the Actor-Actor Network was met with enthusiasm. It allowed users to trace relationships between individuals, making the exploration of social and professional connections more intuitive. As P5 stated, "I am quite enthusiastic about the network views. I think the most logical way to understand people is also through connections."

Do users' perceptions of "interestingness" align with our operational definition? In Figure 1, participants frequently described their findings as both useful and unexpected. In three instances (unexpectedness = 2), participants stumble upon the finding by chance while navigating the portal and found it to be directly relevant to their research interests. The largest group of findings (unexpectedness = 1 and usefulness = 1 or 2) happens when they encountered previously unknown connections between actors and objects, unnoticed production place patterns, or temporal insights enabled by data summarisation, visualisation or semantic linking, they would not otherwise noticed. In four of the six cases (unexpectedness=-1 and usefulness = 1), addressed known gaps in the museum's database. Participants found such insights valuable, as they either helped update existing records or pointed to opportunities for further research to enrich the data. Overall, the responses on the Likert scale revealed that the perceptions of the participants of 'interestingness' were closely aligned with our operational definition. This was reflected in the average score of the two dimensions being more than zero: unexpectedness (0.4±1.063) and usefulness (1.14±0.733).

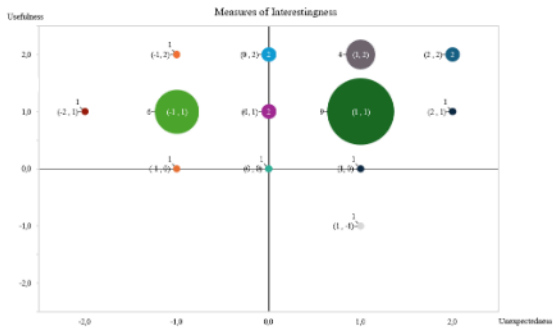

Figure 1: Agreement scores for the 36 findings, plotted across two measures of interestingness: unexpectedness (X-axis) and usefulness (Y-axis).

How can the system be further refined to enhance its effectiveness? : User-study also shed light on improvement areas to better meet user needs. A key issue was the challenge of navigating object origins, as place names are often recorded at highly local levels that users find unfamiliar. This calls for a hierarchical system for browsing places, sortable in alphabetical order, alongside a place perspective that displays related objects, historical events, and connected actors. Participants also recommended clustering objects by internal id, a requirement that surfaced only during the user-study and was previously unknown to the author. Furthermore, adding hyperlinks to all visual summaries would be beneficial, as users often wish to click on points of interest expecting to find more information. Bidirectional relationships should visualise connections from both entity perspectives, which were initially unavailable. For instance, while historical events are linked from an actor's instance page, the reverse link was missing. Additionally, there are conceptual confusions, especially with acquisition dates (date object acquired by the museum) being mistaken for collection dates, underscoring the necessity for clearer labels and informative cues. Participants also suggested features such as a global search to ease exploration. These insights inform strategic enhancements to better support scholarly provenance research.

Overall tool's effectiveness Post-study survey on the tool's evaluation reveals a high demand for semantic exploratory systems in cultural heritage provenance research, as reflected in participant feedback. All participants strongly agreed that PM-Sampo supports exploratory analysis relevant to their research (Q4), and they would consider using PM-Sampo for exploring heritage object provenance (Q1), noting its potential to uncover relationships or insights they might not have discovered otherwise (Q2). Furthermore, 3 participants were open to using PM-Sampo in their research (Q3), while 2 were cautious due to incomplete nature of data. Overall, the user intent indicates the tool's relevance and applicability in supporting knowledge discovery in provenance research.

In conclusion, the user study confirms the effectiveness of PM-Sampo in supporting KD, with participants consistently identifying insights that were both useful and unexpected, aligning with our operationalisation of "interestingness". This is largely enabled by interface features, such as intuitive network graphs, interactive maps, and informative timelines, which allow users to traverse complex relationships and identify patterns previously hidden by exploring explicit, abstract, and implicit semantic associations. Observational and qualitative findings reveal three pathways to KD: (1) curiosity-driven (novel), (2) accidental or serendipitous (novel

and unexpected), and (3) belief-challenging (unexpected), with examples available in the Zenodo repository—highlighting the tool's ability to support both systematic and serendipitous insight.

5 CONCLUSION

In this paper, we address the challenge of enabling knowledge discovery (KD) in the context of provenance research, where data are often incomplete, semantically fragmented, and historically biased. Grounded in the Exploratory Knowledge Discovery (EKD) paradigm, we proposed a novel, domain-centric design approach to support insight generation in Linked Data environments where direct querying is insufficient for addressing complex, interpretive research tasks. The evaluation highlights the general acceptability of the tool for its intended KD purpose and confirms the demand for systems that enable open-ended semantic exploration in cultural heritage research. In addition, the ability of the portal to show information gaps transforms limitations into opportunities for enrichment, strengthening its value in iterative research workflows. This work offers both a methodological contribution to cultural heritage and KD communities and a practical tool to advance the study of provenance in ethically and historically significant contexts.

Our pipeline was designed with modularity in mind to facilitate reusability and reproducibility across domains. Although our design approach proved successful in this domain, it would be valuable to test its applicability in other domains to assess its generalisability. Overall, this work serves as a proof-of-concept demonstrating how EKD can enhance the utility of knowledge graphs with data gaps by adopting a structured and domain-centred approach. The work contributes toward the development of richer, more transparent, and context-sensitive knowledge ecosystems.

Acknowledgement: Thanks to the participants of our user study for their time, feedback, and engagement, which greatly contributed to the evaluation and refinement of our web application.

REFERENCES

[1] Peter Afflerbach. 2001. Verbal reports and protocol analysis. In *Methods of literacy research*. Routledge, 97–114.

[2] Annastiina Ahola, Telma Peura, and Eero Hyvönen. 2025. Using Linked Data for Data Analytic Literary Research: Case BookSampo – Finnish Fiction Literature on the Semantic Web. *Journal of the Association for Information Science and Technology (JASIST)* 76, 7 (2025), 937–958. https://doi.org/10.1002/asi.24984

[3] Mehwish Alam, Aleksey Buzmakov, and Amedeo Napoli. 2018. Exploratory knowledge discovery over Web of Data. *Discrete Applied Mathematics* 249 (2018), 2–17. https://doi.org/10.1016/j.dam.2018.03.041

[4] Aba-Sah Dadzie and Emmanuel Pietriga. 2016. Visualisation of Linked Data – Reprise. *Semantic Web* 8, 1 (2016), 1–21. https://doi.org/10.3233/SW-160249 arXiv:https://journals.sagepub.com/doi/pdf/10.3233/SW-160249

[5] Vania Dimitrova, Lydia Lau, Dhavalkumar Thakker, Fan Yang-Turner, and Dimoklis Despotakis. 2013. Exploring exploratory search: a user study with linked semantic data. In *Proceedings of the 2nd International Workshop on Intelligent Exploration of Semantic Data* (Paris, France) *(IESD '13)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/2462197.2462199

[6] Martin Doerr. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine* 24, 3 (2003).

[7] Usama Fayyad. 1997. Knowledge discovery in databases: An overview. In *Inductive Logic Programming*, Nada Lavrač and Sašo Džeroski (Eds.). Springer, Berlin, Heidelberg, 1–16. https://doi.org/10.1007/3540635149_30

[8] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD-96 Proceedings*.

[9] Dhouha Grissa, Blandine Comte, Mélanie Pétéra, Estelle Pujos-Guillot, and Amedeo Napoli. 2020. A hybrid and exploratory approach to knowledge discovery in metabolomic data. *Discrete Applied Mathematics* 273 (2020), 103–116. https://doi.org/10.1016/j.dam.2018.11.025 Advances in Formal Concept Analysis: Traces of CLA 2016.

[10] Olaf Hartig. 2012. SPARQL for a Web of Linked Data: Semantics and computability. In *Extended Semantic Web Conference*. Springer, 8–23.

[11] Philipp Heim, Thomas Ertl, and Jürgen Ziegler. 2010. Facet Graphs: Complex Semantic Querying Made Easy. In *The Semantic Web: Research and Applications*, Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 288–302.

[12] Eero Hyvönen. 2022. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. *Semantic Web* 14, 4 (2022), 729–744. https://doi.org/10.3233/SW-223034

[13] Eero Hyvönen. 2025. Serendipitous knowledge discovery on the Web of Wisdom based on searching and explaining interesting relations in knowledge graphs. *Journal of Web Semantics* 85 (2025), 100852. https://doi.org/10.1016/j.websem.2024.100852

[14] Eero Hyvönen, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2025. Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland. *Semantic Web* 16, 1 (2025). https://doi.org/10.3233/SW-243683

[15] Esko Ikkala, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2022. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web* 13, 1 (2022), 69–84. https://doi.org/10.3233/SW-210428

[16] Ali Khalili, Pek Van Andel, Peter Van Den Besselaar, and Klaas Andries De Graaf. 2017. Fostering Serendipitous Knowledge Discovery using an Adaptive Multigraph-based Faceted Browser. In *Proceedings of the Knowledge Capture Conference*. ACM, Austin TX USA, 1–4. https://doi.org/10.1145/3148011.3148037

[17] Lynn J. Lohnas. 2025. A Retrieved Context Model of Serial Recall and Free Recall. *Computational Brain & Behavior* 8, 1 (March 2025), 1–35. https://doi.org/10.1007/s42113-024-00221-9

[18] Heikki Rantala, Annastiina Ahola, Esko Ikkala, and Eero Hyvönen. 2023. How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. In *Proceedings of 8th International Workshop on the Visualization and Interaction for Ontologies and Linked Data, Athens, Greece*. CEUR Workshop Proceedings, Vol. 3508. https://ceur-ws.org/Vol-3508/paper3.pdf

[19] Margarete Sandelowski. 2008. Findings. In *The SAGE Encyclopedia of Qualitative Research Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781412963909.n176

[20] Felwine Sarr and Bénédicte Savoy. 2018. *The Restitution of African Cultural Heritage*. Ministère de la Culture.

[21] Sarah Binta Alam Shoilee, Annastiina Ahola, Heikki Rantala, Eero Hyvönen, Victor de Boer, Jacco van Ossenbruggen, and Susan Legene. 2025. Enhancing Provenance Research with Linked Data: A Visual Approach to Knowledge Discovery. In *Proceedings: SemDH 2025 Second International Workshop of Semantic Digital Humanities, co-located with ESWC 2025, Portoroz, Slovenia*. CEUR Workshop Proceedings.

[22] Sarah Binta Alam Shoilee, Annastiina Ahola, Heikki Rantala, Eero Hyvönen, Victor de Boer, Jacco van Ossenbruggen, and Susan Legene. 2025. PM-SAMPO: Semantic Portal for Heritage Object Provenance Research. In *The Semantic Web: ESWC 2025 Satellite Events, Portoroz, Slovenia, June 1 - 5, 2025, Proceedings*. Springer-Verlag.

[23] Sarah Binta Alam Shoilee, Victor de Boer, and Jacco van Oseenbruggen. 2023. Polyvocal Knowledge Modelling for Ethnographic Heritage Object Provenance. In *Knowledge Graphs: Semantics, Machine Learning, and Languages*, Vol. 56. IOS Press, 127.

[24] Abraham Silberschatz and Alexander Tuzhilin. 1995. On subjective measures of interestingness in knowledge discovery.. In *KDD*, Vol. 95. 275–281.

[25] Arthur Tompkins. 2021. *Provenance Research Today*. Lund Humphries.

[26] J. W. Tukey. 1977. *Exploratory Data Analysis*. Pearson.

[27] D. Tunkelang. 2009. *Faceted search*. Morgan & Claypool.

[28] Hannah Turner. 2020. *Cataloguing culture: legacies of colonialism in museum documentation*. UBC Press.

[29] Guillermo Vega-Gorgojo. 2024. LOD4Culture: Easy exploration of cultural heritage linked open data. *Semantic Web* 15, 5 (2024), 1563–1592. https://doi.org/10.3233/SW-233358 arXiv:https://journals.sagepub.com/doi/pdf/10.3233/SW-233358

[30] Jörg Waitelonis and Harald Sack. 2009. Towards Exploratory Video Search Using Linked Data. *In Proceedings of the 2nd IEEE International Workshop on Data Semantics for Multimedia System and Applications (DSMSA2009), in conjunction with IEEE International Symposium on Multimedia (ISM2009)* (Dec. 2009), 540–545.

[31] Dawid Wiśniewski, Jedrzej Potoniec, Agnieszka Ławrynowicz, and C. Maria Keet. 2019. Analysis of Ontology Competency Questions and their formalizations in SPARQL-OWL. *Web Semant.* 59, C (dec 2019), 19 pages. https://doi.org/10.1016/j.websem.2019.100534

[32] Liyang Yu and Liyang Yu. 2011. Follow your nose: a basic semantic web agent. *A Developer's Guide to the Semantic Web* (2011), 533–557.